DOCUMENT RESUME

ED 339 201 FL 019 634

AUTHOR Aghbar, Ali A.; Tang, Huixing
TITLE Partial Credit Scoring of Cloze-Type Items.
PUB DATE 91
NOTE 24p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cloze Procedure; College Students; English (Second Language); *Language Tests; Nouns; *Scoring; *Scoring Formulas; Second Language Instruction; *Test Items; Test Validity; Verbs
IDENTIFIERS *Rasch Partial Credit Model

ABSTRACT

A study was undertaken to develop a partial credit scheme for scoring cloze-type questions on an English collocation test, obtain construct validity evidence for the test and the scoring scheme using the Rasch Partial Credit Model, and compare partial credit scoring with the more commonly used dichotomous scoring with the same test instrument. Subjects were 205 students of English as a Second Language at two colleges in western Pennsylvania. The instrument used tested mastery of verb-noun collocations such as "gain admission." Examinee responses were scored with both the partial credit scale and a conventional dichotomous scale. Statistical analysis of the scoring indicates that the partial credit scoring provides improved item discrimination and external validity, suggesting that it is more desirable both theoretically and psychometrically than the dichotomous method. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Ali A. Aghbar, Indiana Univ. of Pennsylvania
Huixing Tang, University of Pittsburgh
1991 Language Testing Research Colloquium
Educational Testing Service, Princeton, N. J.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Aghbar, A.
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED389201

PARTIAL CREDIT SCORING OF CLOZE-TYPE ITEMS

INTRODUCTION

It has been well accepted in second language acquisition research that acquisition of target language (TL) forms does not take place in an all-or-none fashion; rather, learners follow a systematic developmental route in graduadually approximating the TL forms (Dulay, Burt and Krashen 1982). Constructs measured by language tests such as grammatical ability, listening ability, or reading comprehension are best conceived as continuous and developmental in nature. Yet popular modes of standardized testing, chacterized by dichotomous scoring, are not sensitive to the developmental nature of this approximation process at the item level. By characterizing a response as either right or wrong, the items in such tests dichotomize what is theoretically recognized as a continuum, and do not distinguish among "incorrect" responses which may in fact indicate varying degrees of TL mastery. The widely used cloze tests, for example, are typically scored dichotomously, with the best answer given a score of 1 and all the other answers 0, irrespective of the fact that those uncredited responses may very well indicate varying levels of proficiency. Such a scoring model is lacking in structural fidelity, in the sense that it is not consistent with the recognized behaviorial manifestations of the construct (Messick 1989). Not incidentally, such scoring methods may also lead to undesirable psychometric results.

The purpose of this study is to (a) develop a partial credit scheme for scoring cloze-type questions on an English collocation test (ECT), (b) obtain construct validity evidence for the test and the scoring scheme using the Rasch Partial Credit Model (RPCM), and (c) compare partial credit scoring with dichotomous scoring used with the same test instrument. It is hoped that this study will contribute to the development of scoring models for cloze type questions in language tests which have a more rational theoretical foundation and better psychometric properties.

RATIONALE FOR THE USE OF COLLOCATIONS

There has been a slow but steady recognition of fixed expressions as an important part of language fluency (e.g., Bolinger 1976; Fillmore 1979; Pawley and Syder 1983; Carter and McCarthy 1988). Fixed expressions include collocations, phrases, and sentences whose comprehension and production depend not only on the mastery of constituent lexemes and grammatical rules in them but also on a memory of them as language chunks through previous exposure and experience with the language in communicative situations.

FL019634

BEST COPY AVAILABLE

Native-like fluency depends heavily on the mastery of fixed expressions. Pawley and Syder (1983, p. 205) posit, "The number of memorized complete clauses and sentences known to the mature English speaker is probably many thousands." We believe the command of collocations is no less significant. This can be demonstrated by a quick look at The BBI Combinatory Dictionary of English (Benson, Benson, and Ilson 1986), which in our rough estimation, contains over 60,000 listings of collocations that are used as chunks in both spoken and written English. Here are two sets of collocations from page 223:

sheet	*ship*
bed sheet	build a ship
cotton sheet	refit a ship
percale sheet	launch a ship
double sheet	navigate a ship
fitted sheet	sail a ship
flat sheet	scuttle a ship
king-sized sheet	sink a ship
queen-sized sheet	torpedo a ship
single sheet	abandon ship
twin sheet	jump ship
to change the sheets	raise a sunken ship
to put on clean sheets	load a ship
balance sheet	unload a ship
flow sheet	capital ship
blank sheet	hospital ship
scandal sheet	merchant ship
charge sheet	oceangoing ship
cookie sheet	passenger ship
	rocket ship
	etc., etc.

Although there has been a growing recognition of the role of fixed expressions in language fluency, they have not received the recognition they deserve in language testing. This study explores the possibility of using an English collocation test to assess the mastery of verb-noun collocations such as 'gain admission'.

The English Collocation Test (ECT) has 30 items. Each item consists of one sentence in which a blank represents a deleted main verb. The examinees are expected to fill in each blank with a verb, the choice of which depends on its collocatability with the noun following the blank. Here is an example:

1. My best friend tried to _____ admission to Indiana University of Pennsylvania, but he did not succeed.

Following Sinclair's (1966) view that collocations are best studied in relation to particular registers and situations of use, we asked our subjects to provide a verb that would be appropriate for formal written English in places where they felt there were more than one possibility. (See Appendix A for a complete listing of our test items).

The scoring scheme we used reflects our view that acquisition and use of collocations moves along a continuum from the least semantic approximation to full command of collocations that are idiomatic and appropriate both semantically and by register. Thus, the construct measured by ECT, proficiency in using

English verb-noun collocations, is defined in terms of four ordered levels based on the combinations of the different values of three intertwined facets: *idiomaticity*, *semanticity*, and *register*. The idiomaticity facet has two values: (A) idiomatic and (B) not idiomatic. The semantic facet is construed as having three values: (A) semantic, (B) marginally semantic, and (C) not semantic. Register appropriacy has two values: (A) proper register and (B) not proper register. The combinations of these values define the following ordered levels of the construct: (The examples in the parentheses pertain to the above test item.)

Level 4: idiomatic, semantic, and proper register (e.g. gain)
Level 3: not idiomatic but semantic and proper register (e.g. attain, obtain)
Level 2: not idiomatic, marginally semantic (e.g. earn, take, have), or
idiomatic, semantic but not proper register (e.g. get)
Level 1: not idiomatic, not semantic (e.g. apply, pass, do, live)

The partial credit scoring scheme is strictly derived from the above operationalization of the construct. Specifically, scores are assigned according to the levels of the construct. Since there are four levels, a four point scoring scale (0 to 3) is used with the highest level given a score of 3 and the lowest level assigned a score of 0. Thus, one notable feature of this scoring scheme is that it is consistent with the hypothesized set of ordered categories of the construct behavior.

METHOD

1. Subjects.

The test was administered to 205 ESL students at two colleges in Western Pennsylvania. Of these, 122 were new freshman students who took the test as part of the English placement battery at Indiana University of Pennsylvania in Fall 1990 and Spring 1991. Twenty-eignt were in their sophomore year at IUP. The remaining 55 students were in preparatory intensive English programs at IUP and another college.

2. Materials

All the subjects took a 30-item collocational test in the format discussed above. The new freshman students took this test as part of the English Placement Test, which additionally includes a composition written in 35 minutes and scored holistically, a 50- item multiple-choice cloze test, and a read-comp test, where the students read a paragraph in 3 minutes and rewrite it from memory in 7 minutes. The read-comp is scored holistically using a rubric as the guide. In addition, TOEFL scores were obtained for the new freshman students at IUP.

To compile a list of the most common collocations produced by educated native speakers of English, we asked the following native speaker groups to take the test:

27 instructors of English at IUP
21 students majoring in English education
23 sophomores in majors other than English

All the responses were entered into a Dbase III Plus file and then exported into an SPSS system file for further processing.

3. Scoring Procedures

A program for scoring was created by the researchers, one of whom is an expert in the study of collocations. The raters assigned scores to responses together using the native speaker responses as a general guide. In cases where they differed in judgments, they discussed the scores and the rationale until consensus was reached. The raters reviewed score assignments after two weeks and ironed out any perceived inconsistencies. Once the program was ready, the scoring was done using the recode function in SPSSX to translate responses in words into score codes.

4. Analysis Procedures

For purpose of analysis, examinee responses are scored on both the partial credit scale described earlier and a dichotomous scale following the conventional pocedure of assigning a score of 1 to the best answer and a score of 0 to all the other answers. The polytomously scored data set were analyzed using the Rasch Partial Credit Model proposed by Master (1982). This model is appropriate for data coming from "an observation format which requires the prior identification of several ordered levels of performance on each item and thereby awards partial credit for partial success on items" (Master 1982, p.150). It differs from Andrich's rating scale models in that the distances between the levels or steps are free to vary both within an item and across items and the number of steps for each item is also free to vary across items. It also differs from the graded response model proposed by Samejima(1969) in that the estimates of the steps in RPCM do not necessarily have to be in the assumed order. An advantage of this property of RPCM is that "it is possible for the data not to accord with the model, that is, it is possible for the data to refute the model with respect to the criterion of order on the latent continuum" (Andrich 1985, p.253). Since the construct and the scoring scheme we are trying to validate involve ordered levels, the distances between which are presumably unequal both within and between items and since the hypothesized order of the levels is yet to be verified empirically, we considered RPCM most relevant for the present study. Use of this model will provide some evidence useful for addressing the following interrelated validity questions:

1. To what extent is the theoretically specified order of the categories supported empirically?

2. To what extent do the items fit the model and can therefore be used as valid measurements of the construct?

3. To what extent are the students measured validly, or how credible is this measurement instrument for the population of ESL students that the ECT is intended for?

Prior to performing Rasch analysis, we felt it was necessary to check the assumptions of (a) unidimensionality, (b) equal discrimination, and (c) minimal guessing. The latter two assumptions are required since RPCM is a one-parameter model. To check the unidimensionality assumption, principal factor analysis was run on the polytomous score data using the 6.06 release of the Statistical Analysis System, "Proc Factor" (SAS Institute, 1988: 449-492). The results indicate that one factor would adequately represent the data since there is only one eigenvalue greater than the value of 1 and the scree plot (see Appendix I) of the eigenvalues against measures drops sharply and levels off after the first eigenvalue. Thus the unidimensionality assumption is satisfied. As a rough check of equal discrimination assumption, we examined the distribution of the polybiserial correlations computed for each item, following the suggestion offered by Hambleton (1985). The results indicate that the distribution is reasonably homogeneous with a mean of 0.582, a standard deviation of 0.0699 and a range of 0.2975 (see Appendix II). Thus, the equal discrimination assumption also seems to be tenable. As for the assumption of minimal guessing, we expect the guessing effect to be minimal since the test is in the open- ended format and no response cues are provided.

According to our original design we were to analyze the dichotomously scored data set using a dichotomous Rasch model so that the results could be directly compared with those obtained from the polytomous model. The examination of the underlying assumptions reveals that the dichotomous data do not fit the model. In particular, the factor analysis shows that there are three eigen values greater than the value of 1 (see Appendix I) with the first eigen value accounting for no more than 8% of the variance and the distribution of point-biserial correlations has a standard deviation of 0.119 and a range of 0.5381, both of which are too big for the distribution to be considered reasonably homogeneous (see appendix II.) Thus neither the unidimentionality assumption nor the equal discrimination assumption is tenable. With the assumptions not being satisfied, we find it imprudent to proceed with Rasch analysis on the dichotomously scored data set since the estimates from such an analysis are likely to be biased. We have decided, instead, to use the conventional methods to compare the two scoring schemes and perform Rasch analysis only on the partial credit score data.

The RPCM analysis was done using version 2.02 of the Bigsteps software package (Wright and Linacre, 1991) on IBM PC. It uses PROX to make initial, rough estimates and then the UCON method to calculate more exact estimates, standard errors, and fit statistics (Wright and Linacre, 1991). Other statistical analysis was done using SAS release 6.06 and SPSSX release 4.0. For ease of reference, each item is labeled by the collocational phrase tested by that item.

RESULTS & DISCUSSION

Examining step estimates

Since one of the major concerns in this study is the extent to which the order of the levels defined in our operationalization of the construct is supported empirically, we start by examining the step estimates of RPCM for the data. As the examinee responses are scored on a four point scale, The RPCM analysis of

the data provides three step estimates for each item, d_{i1},d_{i2} and d_{i3}, if there are responses in each of the categories. These estimates govern the probability of scoring 1 rather than 0, 2 rather than 1, and 3 rather than 2 on each item. Table 1 provides these estimates along with their calibration errors. One may note that for items 13 and 15, there are only two step estimates because in each item, there are no examinees who scored in the third category.

Table 1. Step estimates and calibration errors

	ITEM STEP ESTIMATES			ESTIMATION ERRORS		
NAME	d_1	d_2	d_3	s_1	s_2	s_3
1	-1.83	1.41	.42	.23	.17	.20
2	.64	.48	-1.12	.23	.20	.18
3	.21	.37	-.58	.19	.18	.17
4	-1.27	1.59	-.32	.18	.19	.22
5	.13	.70	-.83	.19	.18	.17
6	-2.04	1.80	.25	.20	.20	.27
7	-.87	-.02	.89	.21	.16	.19
8	-.57	-.53	1.11	.18	.16	.23
9	-.33	1.93	-1.60	.21	.18	.17
10	-1.22	.83	.39	.18	.18	.24
11	-.77	1.34	-.57	.16	.21	.24
12	-.19	1.13	-.94	.16	.25	.30
13	-.16	.16		.18	.16	
14	.83	-2.14	1.31	.18	.17	.21
15	.34	-.34		.16	.20	
16	-.98	1.23	-.25	.20	.17	.18
17	-.31	.44	-.13	.17	.18	.20
18	-.31	.65	-.34	.17	.18	.19
19	1.13	-1.35	.22	.17	.18	.22
20	.20	.33	-.53	.17	.18	.20
21	-.84	.93	-.09	.16	.19	.23
22	-.14	.31	-.17	.17	.17	.19
23	-1.40	2.03	-.64	.18	.20	.23
24	.04	.57	-.61	.17	.23	.27
25	-.54	1.57	-1.03	.16	.26	.30
26	-.73	1.97	-1.24	.17	.19	.20
27	-2.27	1.01	1.26	.24	.17	.26
28	-.28	-.18	.47	.18	.17	.20
29	-.93	.54	.39	.16	.19	.26
30	-1.60	.73	.86	.18	.18	.29

The step estimates for each item define a unique set of probability curves for the performance levels measured by that item. Figure 1 shows the curves defined by three step estimates for item 7 'impose tax'. The estimates d_{71}=-.87, d_{72}=-.02 and d_{73}=.89 are located at the intersection of successive category probability curves. These curves show that the most probable score is 0 for students with ability estimates below -.87, 1 for students with estimates between -.87 and -.02, 2 for students with estimates between -.02 and .89, and 3 for those with estimates greater than .89.

Fig 1. RESPONSE CATEGORY PROBABILITY CURVES FOR ITEM 7

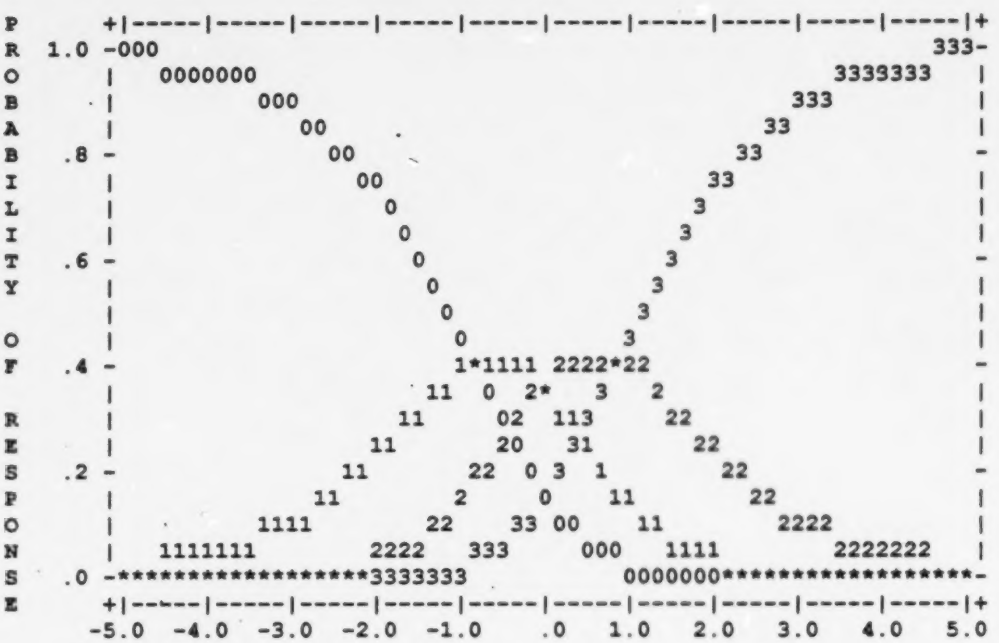
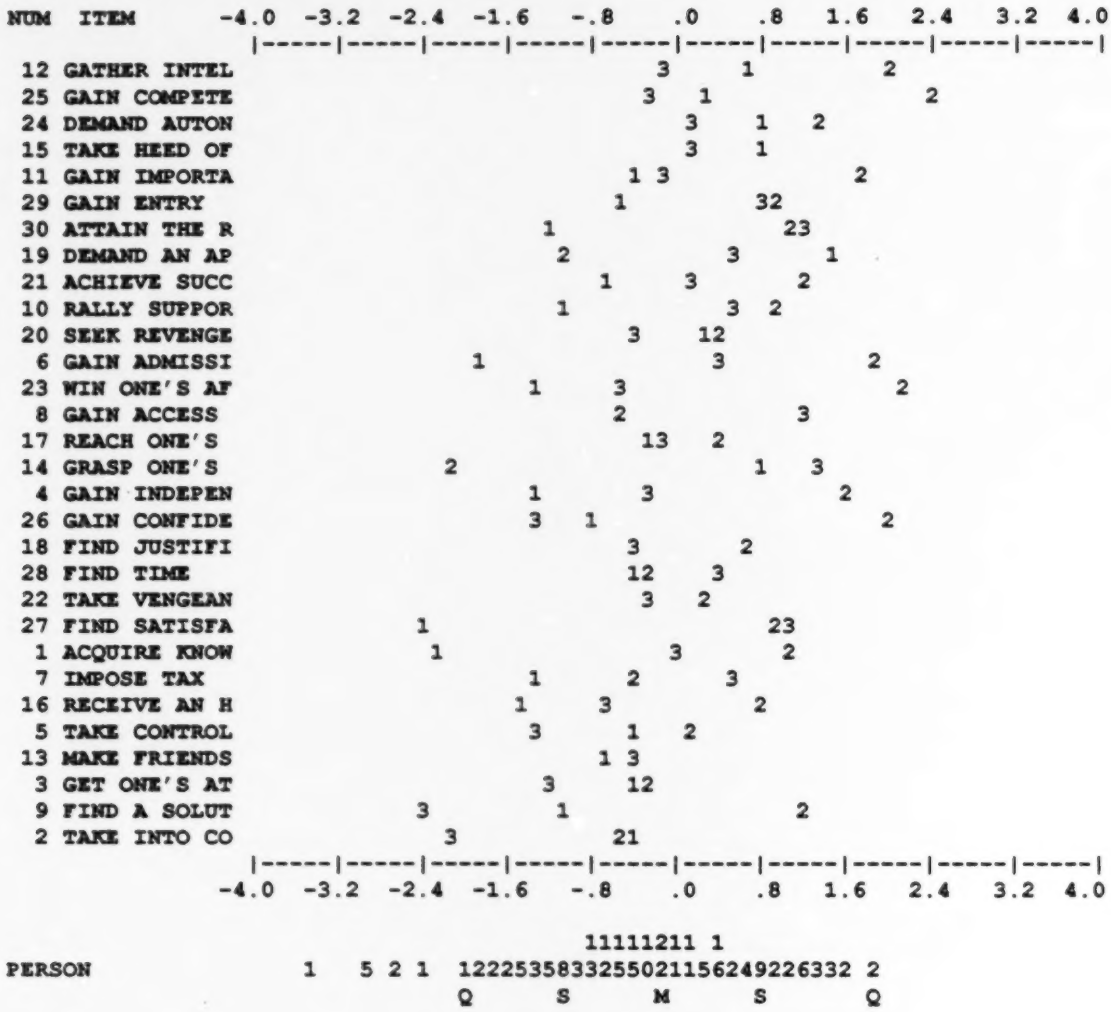


Fig 2 is the plot of the step estimates for each item, listed in descending order of difficulty. So far as step ordering is concerned, there are basically three groups of items. The first group includes six items (7, 8, 13, 27, 28. 30), in which step ordering matches our prior theoretical specification. The second group includes 10 items (1, 4, 6, 10, 11, 16, 17, 21, 23, 29), in which both the second and third step are more difficult than step 1, but step 2 has higher estimates than step 3. This means that for these items, step 2 responses are never the most likely responses. Students are expected, under the model, to be less likely to complete the second step than the third step once they complete the first step. Take for example item 1, 'acquire knowledge'. For this item, level 4 responses (for completing step 3) are 'acquire' and 'gain'. Level 3 responses include 'attain','achieve','enhance','master'. And level 2 responses include 'have','get','take'. It is not difficult to see that while level 3 and level 4 responses are more appropriate than level 2 responses, the difference between level 3 and level 4 in this case may be too fine for this group of students. They may have learnt the use of 'acquire' or 'gain' as easily, if not more so, as those level 3 expressions. Therefore they are more likely to respond with level 4 expressions than with level 3 expressions once they are beyond level 2 since the former are more idiomatic than the latter.

The third group includes 11 items (3, 5, 9, 12, 15, 18, 20, 22, 24, 25, 2[illegible]. In these items, which mostly involve common verbs such as 'get','take' and 'find' as the best response, both step 1 and step 2

Fig 2. STEP DIFFICULTY CALIBRATIONS



estimates are higher than step 3 estimates, suggesting that for these items, level 2 and level 3 responses are less likely than level 4 responses once students reach beyond level 1. Take item 5 'get someone's attention'. Level 4 responses are 'get' and 'attract'. Level 3 responses include 'gain','catch','draw', 'capture'. And level 2 responses include 'have','take','occupy', 'collect'. It seems clear that some or most of the level 3 and level 2 words are somewhat more 'formal' than level 4 words, and it may be reasonable to infer that these students have acquired those idiomatic expressions (particularly 'get') earlier than those less idiomatic but semantically appropriate or marginally appropriate expressions. Therefore, level 4 responses are more probable once they reach beyond level 1 ability.

The different orderings of the step estimates have different implications for our prior specification of step order. The first group of items fits our specification. If we want to maintain our defined construct, only this

group of items should be retained. The order in the second group suggests that the distinction between level 3 and 4 is too fine or that level 4 is not really more difficult than level 3 for this group of students so that the two highest levels are best considered one level. The order in the third group suggests that two levels would be adequate, level 1 vs. level 4, in which case the facet of semantic appropriacy seems unnecessary. Either the students do not know the answer, or, if they do, they are more likely to know the idiomatic answer.

Fig 2 also shows that the distances between step estimates are quite varied from item to item. For some items the step estimates cover a relatively wide range of difficulty. For others, the step estimates are quite close to each other. Take item 29 and 30. The distance between step 2 and 3 for both items is within 0.15. Such closeness of the distance between steps also raises the question as to whether the distinction between these categories is useful.

Examining Item Fit Statistics

Bigsteps computes two fit statistics for each item: infit and outfit, which indicate the overall fit of the item to the model. Both infit and outfit are standardized (with mean 0 and variance 1) mean square fit statistics. While infit is information-weighted and is more sensitive to unexpected behavior affecting responses to items near the person's ability level, outfit is more sensitive to unexpected behavior by persons on items far from their ability (Wright and Linacre, 1991). Generally, fit values greater than 2 are considered 'positive misfits', indicating noise, and fit values less than -2 are considered 'negative misfits', suggesting item dependency. Item fit statistics are an important source of validity evidence. To the extent an item fits or misfits the model, its use as a measurement of the construct is validated or invalidated.

Table 2 provides summary statistics for all the items, which shows that the mean for both the infit and outfit is close to 0 and the item separation reliability, the ratio of item variance adjusted for calibration error over total item variance, is .96. The latter index indicates the reliability with which the sample separates these items (Wright and Masters, 1982).

Table 2 SUMMARY OF 30 CALIBRATED (NON-EXTREME) ITEMS

	SCORE	SAMPLE	CALIBRTN	ERROR	MNSQ	INFIT	MNSQ	OUTFIT
MEAN	246.3	205.0	.00	.09	1.0	.1	1.0	.1
S.D.	74.6	.0	.44	.01	.1	1.5	.2	1.5
RMSE	.09	ADJ.S.D.	.43	ITEM SEP	5.02	ITEM SEP REL.		.96

Table 3. ITEM FIT STATISTICS -- ENTRY ORDER

NUM	MNSQ	INFIT	MNSQ	OUTFT	PTBIS	NAME
1	1.04	.4	1.10	.9	.51	ACQUIRE KNOWLEDGE
2	1.31	2.6	1.22	1.0	.48	TAKE INTO CONSIDERATION
3	1.18	2.0	1.51	3.0	.50	GET SOMEONE'S ATTENTION
4	.89	-1.2	.91	-.8	.61	GAIN INDEPENDENCE
5	1.05	.6	1.09	.6	.55	TAKE CONTROL OF
6	.78	-2.0	.79	-1.8	.62	GAIN ADMISSION
7	1.05	.6	1.08	.9	.54	IMPOSE TAX
8	.94	-.7	.92	-.7	.59	GAIN ACCESS
9	1.22	2.3	1.64	3.5	.46	FIND A SOLUTION
10	.94	-.6	.92	-.8	.57	RALLY SUPPORT
11	1.06	.6	1.08	.6	.50	GAIN IMPORTANCE
12	.99	.0	1.05	.3	.48	GATHER INTELLIGENCE
13	1.10	1.3	1.12	1.2	.43	MAKE FRIENDS
14	1.36	3.7	1.39	2.4	.36	GRASP ONE'S HAND
15	.91	-1.1	.78	-1.4	.54	TAKE HEED OF
16	.85	-1.8	.82	-1.8	.65	RECEIVE AN HONOR
17	1.08	.9	.95	-.4	.54	REACH SOMEONE'S IDEALS
18	1.09	1.0	1.02	.2	.53	FIND JUSTIFICATION
19	.80	-2.4	1.00	.1	.67	DEMAND AN APOLOGY
20	1.09	1.0	1.22	1.4	.53	SEEK REVENGE
21	.89	-1.1	.82	-1.6	.61	ACHIEVE SUCCESS
22	1.11	1.2	1.17	1.4	.52	TAKE VENGEANCE
23	.89	-1.1	.80	-1.8	.60	WIN SOMEONE'S AFFECTION
24	1.02	.2	.94	-.2	.52	DEMAND AUTONOMY
25	1.02	.2	.95	-.2	.49	GAIN COMPETENCE
26	.84	-1.8	.80	-1.8	.66	GAIN CONFIDENCE
27	1.06	.6	1.16	1.4	.43	FIND SATISFACTION
28	1.05	.6	1.18	1.5	.54	FIND TIME
29	.75	-2.8	.73	-2.7	.68	GAIN ENTRY
30	.96	-.4	.91	-.9	.55	ATTAIN THE RANK

Table 3 provides fit statistics for each item. We note that there are four items (2,3,9,14) with infit values greater than or equal to 2 and three of them also have outfit values greater than 2. Large positive fit values are generally caused by some examinees scoring higher than expected by the model and some others scoring lower than expected. Take item 2, 'take into consideration', which has a positive infit value of 2.6. To examine the response frequencies of different ability groups for different score categories, we divided subjects into four ability groups, group 1 through group 4, according to their ability measures. Group 1 refers to the least proficient group and group 4 refers to the most proficient group. We constructed a frequency table which lists the frequencies as well as percentages for each group and for each score category. Fig 3 below is the SAS table for Item 2. We note that for this item, the percentages of the two low proficiency groups scoring 3 are surprisingly high, 57.14 percent for the group 2 and 36.36 percent of for group 1. On the other hand, 14 percent of the group 3 scored 0. This is quite high considering that only 16.33 percent of group 2 scored 0. This kind of unexpected frequency pattern is the cause of the misfit of the item.

Fig 3. TABLE OF GROUP BY SCORE CATEGORY

LEVEL I2

Frequency / Row Pct	0	1	2	3	Total
GROUP1	25	8	2	20	55
	45.45	14.55	3.64	36.36	
GROUP2	8	9	4	28	49
	16.33	18.37	8.16	57.14	
GROUP3	7	2	8	33	50
	14.00	4.00	16.00	66.00	
GROUP4	0	0	4	47	51
	0.00	0.00	7.84	92.16	
Total	40	19	18	128	205

Three of the four positive misfitted items involve the use of common verbs ('take','get','find') as the best answer. These verbs are presumably among the early acquired verbs. Therefore, they are relatively easy for the low proficiency groups and their chances of scoring 3 are high. On the other hand, for some high ability students, their concern for register may override their concern for idiomaticity and this may prompt them to avoid using those common verbs and opt, instead, for those more formal but less idiomatic expressions, resulting in their scoring lower than expected on these easy items. This could also be the cause of misfitting.

Also to be noted from table 3 is that there are three items which have fit values less than -.2. The negative misfits are generally caused by items' discrimination being higher than expected by the model. While this would normally be considered desirable, it deviates away from model expectation and will lead to measurement inconsistencies from the model point of view.

Examining person misfits

Table 4. SUMMARY OF 205 MEASURED (NON-EXTREME) PERSONS

	COUNT	TEST	MEASURE	ERROR	MNSQ	INFIT	MNSQ	OUTFIT
MEAN	36.0	30.0	-.35	.24	1.1	.1	1.0	.0
S.D.	18.1	.0	.95	.09	.4	1.2	.4	1.2

RMSE .25 ADJ.S.D. .92 PERSON SEP 3.62 PERSON SEP REL. .93

Person fit statistics indicate the credability of the measurement instrument. To the extent the test fits a person, the measurement of that person with that test is validated. Table 4 gives the mean fit values for the sample, with infit equal to .1 and outfit equal to 0, and a person separation reliability index, the ratio of person variance adjusted for calibration error over total person variance, of .93. The latter index indicates the reliability with which the test discriminates among persons.

Among the 205 subjects there are fifteen of them with either the infit or outfit or both greater than or equal to 2, around 7 percent of the total sample. Of these misfitted persons, over half of them are precollege ESL students with relatively low ability measures. Their responses and the standarized score residuals are provided in Fig 4. The items have been reordered so that they are listed in ascending order of difficulty. The residuals are given in stardard deviation units, with expectation 0 and variance 1. Negative residuals indicate that the examinee's score on an item is lower than expected while positive residuals indicate that the examinee's score is higher than expected. Usually, a residual value greater than 2 or less than -2 is considered significant.

Student 908 has the largest infit and outfit values (4.23 and 4.49, respectively). His response pattern is quite unusual in that he missed four of the five easiest items (item 2, 3, 13, 5) but scored 3 on some of the difficult items (Item 21,19, 29,15). Student 928 has the lowest measure, who missed all the items except item 1 which he scored surprisingly higher than is expected, 5 standard deviation units away from the expected value 0. Most other students listed in the table have low ability measure and their misfits are mostly caused by scoring some items higher than expected given their overall performance. This may be partly due to guessing and partly due to some items being unusually easy for this group of students.

From Table 4 we may see that the mean measure of the whole group is -.35, suggesting that the test as a whole is difficult. This is even more so for the pre-college ESL students, which might be one of the reasons why an unusually large proportion of them are misfitted. This raises the question as to whether ECT can be validly used for ESL students at low proficiency levels, or whether the distnctions made between the four levels of performance are relevant for this group of students considering their quite limited proficiency.

Comparing partial credit scoring with dichotomous scoring

As mentioned earlier, when the same responses are scored on a dichotomous scale, the factor structure and the distribution of the discrimination indices changed. The unidimensionality assumption no longer holds, presumably due to construct-irrelevant variances. The dsicrimination indices decreased, with a

Fig 4. Misfitted persons from the pre-college group

```
------------------------OUTFIT-----INFIT------MEASURE-----------

ITEM        2 9 3 1 5 1 7 1 2 2 2 1 2 4 1 1 8 2 6 2 1 2 1 3 2 1 1 2 2 1
                  3   6     7 2 8 8 6   4 7   3   0 0 1 9 0 9 1 5 4 5 2

908                           4.23       4.49          .00
RESPONSE:1  0 3 0 0 0 3 3 1 2 0 1 3 3 0 2 0 0 3 0 3 0 3 3 1 3 0 3 0 1 2
RESIDUAL:  -3 0-2-1-2 1 1 0 0-1 0 1 1-1 0-1-1 2-1 1-1 1 1 0 2-1 1 0 0 1

910                           2.48       2.32          .15
RESPONSE:1  1 3 3 1 0 1 0 1 2 2 3 3 3 0 2 3 2 1 0 0 1 2 2 2 3 0 3 0 3 0
RESIDUAL:  -2 0 0 0-2 0-2 0 0 0 1 1 1-1 0 1 0 0-1-1 0 0 0 0 1-1 1 0 2 0

922                           2.48       2.53         -.66
RESPONSE:1  3 3 0 0 0 0 2 0 3 2 0 0 0 1 3 0 3 0 0 2 0 0 0 1 0 3 0 0 0 0
RESIDUAL:   0 1-1-1-1-1 0-1 3 1 0 0 0 0 0 1 0 2-1-1 1-1 0 0 0 0 3 0 0 0 0

716                           2.06        .53        -3.10
RESPONSE:1  0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0
RESIDUAL:   0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 5 0 0 7 0 0 0 0 0 0 0

724                           1.83       2.04        -1.22
RESPONSE:1  0 3 0 3 3 1 0 0 3 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1
RESIDUAL:   0 2 0 1 2 0-1-1 3 0 0 0 0 0 0 0 0 0-1 0 0 0 0 0 0 0 0 0 1 2

915                           1.64       2.40        -1.09
RESPONSE:1  3 3 0 0 0 0 1 1 0 0 0 0 0 3 3 0 3 0 0 0 1 0 0 0 0 0 0 0 0 0
RESIDUAL:   1 1 0 0 0-1 0 0-1 0 0 0 0 3 2 0 2-1-1 0 0 0 0-1 0 0 0 0 0 0

919                           1.25       2.30        -1.15
RESPONSE:1  3 3 0 0 0 0 0 1 0 3 1 0 0 0 3 0 2 0 0 0 1 0 0 0 0 0 0 0 0 0
RESIDUAL:   1 2 0 0 0-1-1 0-1 3 0 0 0 0 2 0 1-1-1 0 0 0 0-1 0 0 0 0 0 0

928                            .51       2.33        -3.10
RESPONSE:1  0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
RESIDUAL:   0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

mean drop of .0885 (t=6.16, p=0.0000). The decrease of discrimination is largely a result of merging level 1, level 2 and level 3 responses into one level, thereby losing a great deal of discriminative information of these levels. Take Item 30,'ATTAIN THE RANK'. The poly-biserial correlation (for polytomous scoring) for this item is 0.58 and the point-biserial correlation (for dichotomous scoring) is only 0.25, a dramatic drop of .33. Table 5 provides the response frequencies and percentages for the four ability groups for both polytomous and dichotomous scoring. We note from the table that when the responses are scored polytomously, the percentage of students scoring the second highest score goes increasingly high from group 1 to group 4 and percentage of students scoring 1, the second lowest score, goes increasingly low from group 2 to group 4. This suggests that both of these score categories are discriminating. When the responses are scored dichotomously, the discrimination of these two categories disappear so that the percentage of students scoring 0 is high for all the four groups and over 80% of the group 4 students seem to have the same proficiency as the group 1 students. The loss of information due to dichotomous scoring is quite obvious.

Table 5. Response frequencies and percentages for item 30

		Polytomous				dichotomous		
GROUP		0	1	2	3	0	1	Total
GROUP1	freq	40	14	1	0	55	0	55
	%	72.73	25.45	1.82	0.00	100.00	0.00	
GROUP2	freq	6	37	5	1	48	1	49
	%	12.24	75.51	10.20	2.04	97.96	2.04	
GROUP3	freq	4	32	10	4	46	4	50
	%	8.00	64.00	20.00	8.00	92.00	8.00	
GROUP4	freq	5	19	17	10	41	10	51
	%	9.80	37.25	33.33	19.61	80.39	19.61	
Total		55	102	33	15	190	15	205

As mentioned earlier, ECT was administered together with some other tests for the ESL college students in the sample. Since these tests measure similar, though not identical, language constructs, we may expect positive correlation between ECT and these tests. Fig 5 provides the correlation coefficients between the ECT total score from polytomous scoring (Total4) and that from dichotomous scoring (Total2) with the other test scores including TOEFL. Even though none of the correlation coefficients exceed .80 due to the limited range of the subgroup, Total4 is consistantly higher than Total2, suggesting better discrimination and higher external validity.

Fig 5. Correlations of polytomous scores and dichotomous scores with external measures

	TOEFL	COMP	CLOZE	READCOMP
TOTAL4	.7841	.5385	.5766	.6844
TOTAL2	.7298	.4713	.5049	.6184

CONCLUDING REMARKS

In this study, we developed a partial credit scoring scheme for an English collocation test. One notable feature of this scoring scheme is that it is derived from the operationalization of the construct measured by the test and is, therefore, consistent with the behavioral manifestations of the construct. In particular, the ordinal score scale corresponds to the ordered levels of the construct behavior. We then used RPCM to obtain evidence to validate the scoring scheme as well as the construct. The RPCM analyses provide for us relevant information to address the validity questions we have raised and to evaluate the adequacy as well as appropriacy of the scoring scheme.

With regard to the first question concerning the validity of the defined order of levels, the evidence is inconclusive. Whether to redefine the number and the order of levels so as to accommodate more data or to maintain the same definition and revise and select items so as to accommodate the defined order remains a question, the answer to which would depend on further theoretical speculation on the construct, the consideration of the purpose of the test, and more empirical data.

As regards the second question concerning item fit, the results indicate most items fit reasonably well. We noted that some misfitting items involve the use of common verbs such as 'get','take' as the best response and do not discriminate as well as the other items in the set.

The results pertinent to the third question regarding person fit seems to suggest that while the test fits the sample as a whole reasonably well with a person saperation reliability of .93, the proportion of misfitting persons in the pre-college ESL group is quite high and this raises questions as to whether the test can be validly used for low proficiency groups and, in particular, whether the number as well as the order of the levels defined is relevant for this group of students.

The comparison of partial credit scoring and dichotomous scoring shows the improved discrimination as well as better external validity with partial credit scoring. This suggests that the polytomous scoring method is not only more rational theoretically, but also more desirable psychometrically than the dichotomous method.

We believe that RPCM as well as conventional statistical analyses have provided valuable information that will be useful in a similar testing attempt in the future. After all, we had devised a test that had not been tried before and whose construct had not been tested. Moreover, the open-choice nature of the test gave us little control over the specific responses we could elicit form the subjects in relation to the construct. We believe this is, in a way, an advantage since, compared with a multiple-choice test, the open-choice test is more reflective of what the students might choose on their own in natural communication; a preliminary pilot study we have done indicates that low-proficiency students are much more likely to choose an appropriate answer in a multiple-choice test even when over ten choices are given for an item.

Despite some of their advantages, open-choice tests are known to be time-consuming and have a lower inter-scorer reliability than MC tests especially when impromptu judgments have to be made. We were able to partially overcome these problems by developing a scoring scheme based on the operationalization of the construct and using a database package for data entry and using the SPSSx for scoring. We feel that the use of fast, efficient computers and spell checkers holds good potentials for open-choice testing. For example, one might devise a program where the students type in their responses on a computer and are given feedback where they have spelling problems. Then, their responses can be evaluated using a scoring program like the one we created. Another program can be created to flag new responses that have not been assessed previously.

REFERENCES

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In: S. E. Embretson (ed). *Test design: Development in psychology and psychometrics*, 245-275. New York: Academic Press, Inc.

Benson, M., Benson, E., and Ilson, Robert (Compilers). (1986). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Philadelphia: John Benjamins Publishing Company.

Bolinger, D. (1976). Meaning and Memory. *Forum Linguisticum 1(1)*: 1-14.

Carter, R. and McCarthy, M. (1988). *Vocabulary and Language Teaching*. New York: Longman.

Dulay, H., Burt, M. & Krashen, S. (1982). *Language two*. New York: Oxford University Press.

Fillmore, Charles J. (1979). On Fluency. In: C. Fillmore, D. & Kempler, W. Wang (eds.). *Individual Differences in Language Ability and Language Behavior*. NY: Academic Press.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Messick, S. (1989). Validity. In: R. Linn (ed). *Educational measurement*. New York: Macmillan Publishing Company.

Pawley, A. and Syder, H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Richard, J. C. and Schmidt, R. W. *Language and Communication:191-225*. Longman.

Sas Institute (1988). *SAS/STAT user's guide, release 6.03 edition*. Cary, NC: SAS Institute Inc.

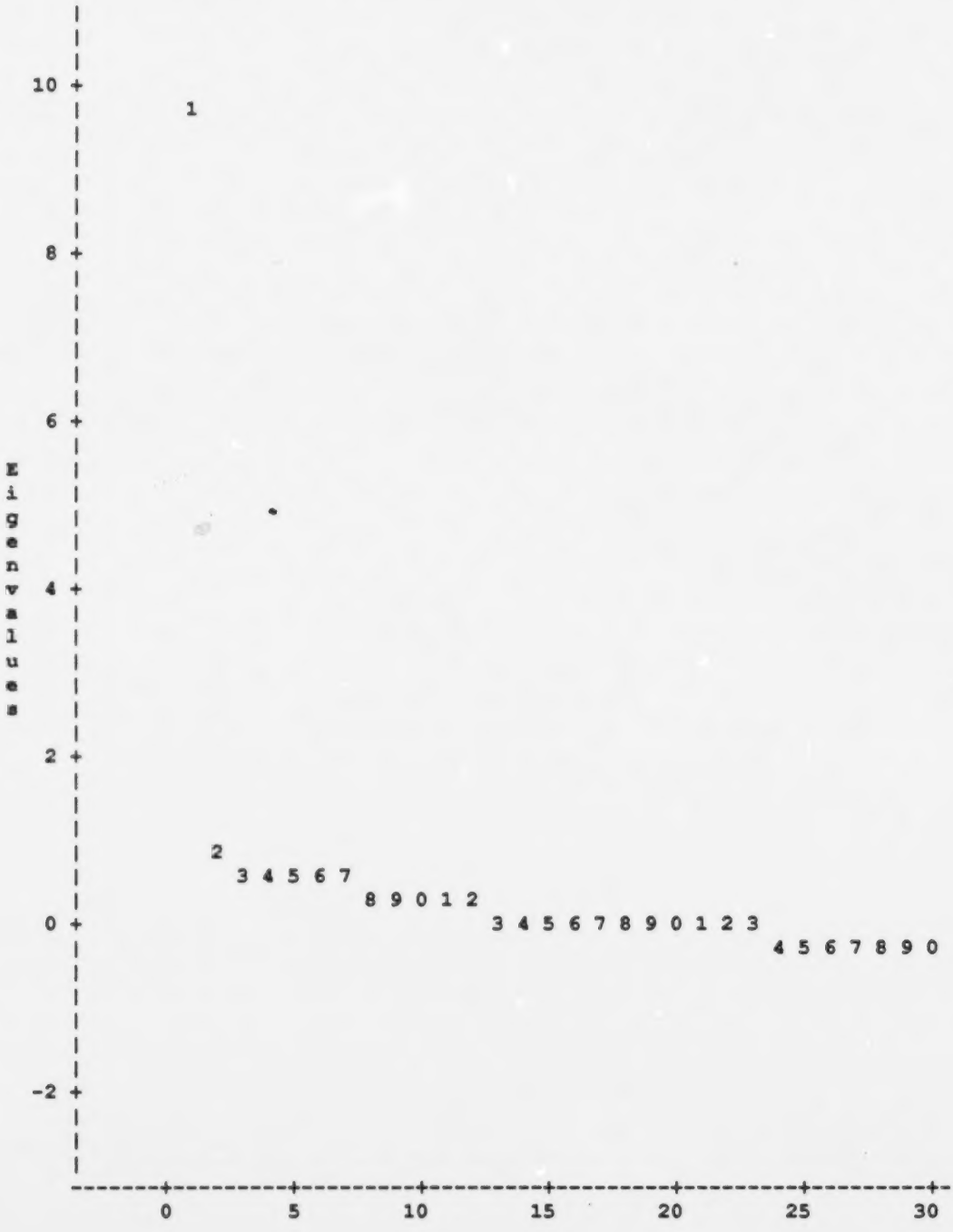
Sinclair, H. McH. 1966. Beginning the study of lexis. In Bazell, C. E., Catford, J. C., Halliday, M. A. K., and Robins, R. H. (eds.) *In Memory of J. R. Firth:410-430*. New York: Longman.

Wright, B. D. & Linacre, J. M. (1991). *A User's Guide to Bigsteps*. Chicago: MESA Press.

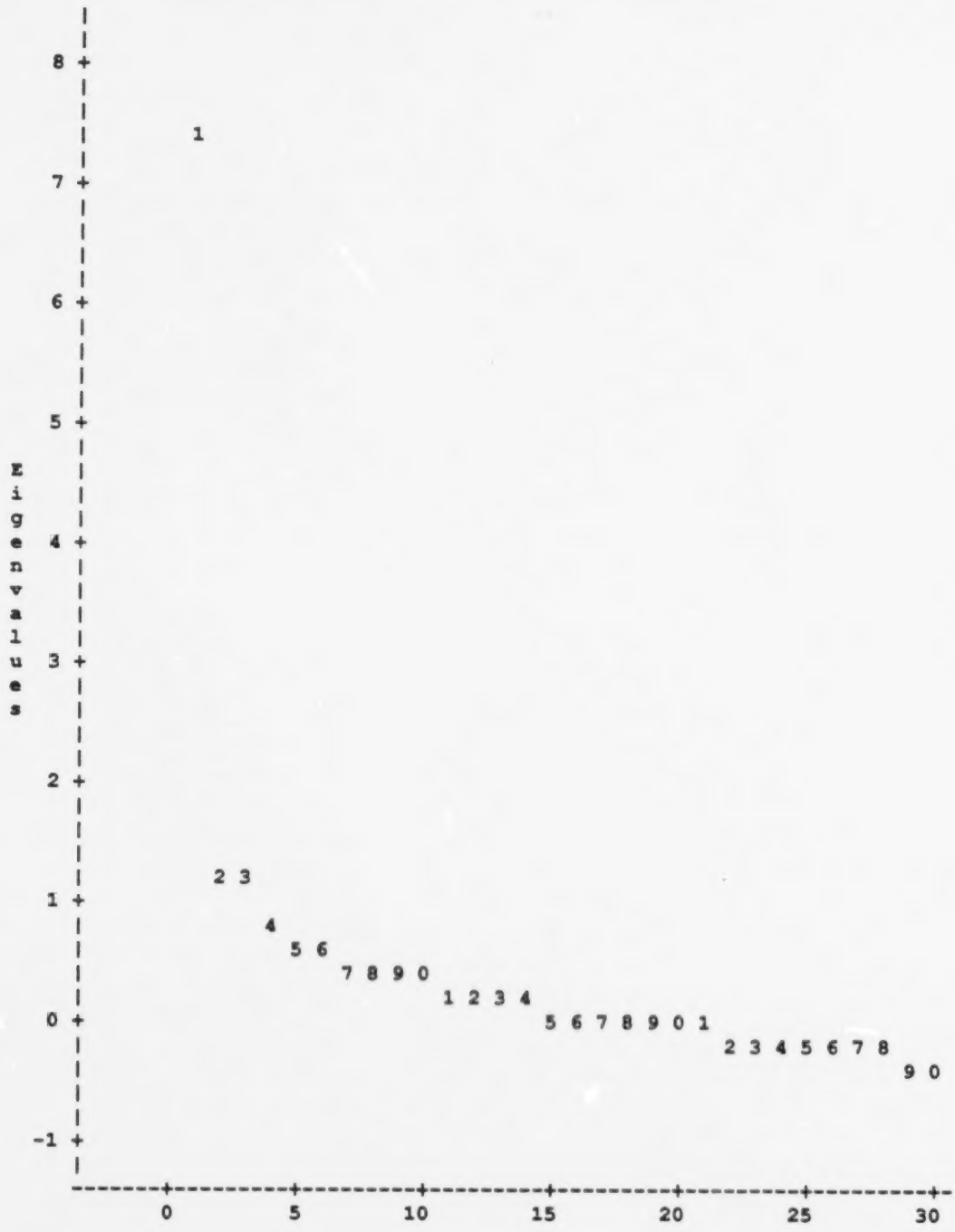
Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press.

APPENDIX I

Scree Plot of Eigenvalues for the polytomous data



Scree Plot of Eigenvalues for dichotomous data



Appendix II. The distributions of the poly-biserial correlations and the point-biserial correlations

Histogram of the polybiserial correlations for the polytomous score data

Midpoint	Count	
0.40	1	*
0.44	0	
0.48	2	**
0.52	4	****
0.56	10	**********
0.60	4	****
0.64	5	*****
0.68	3	***
0.72	1	*

Histogram of point-biserial correlations for the dichotomous score data

Midpoint	Count	
0.15	1	*
0.20	0	
0.25	1	*
0.30	0	
0.35	1	*
0.40	2	**
0.45	5	*****
0.50	9	*********
0.55	2	**
0.60	5	*****
0.65	4	****

Appendix III. English Collocation Test

In each of the following sentences a verb is missing. Supply the verb that you think will best fit each sentence. Imagine these sentences are from formal written texts; therefore, when more than one verb is possible, choose the verb that is appropriate for formal written contexts. Please use only one verb in each blank. Do try to respond to every item because partial credit will be given to answers that are partially correct.

1. This is a good opportunity for you to _____ knowledge in your field of study.

2. I will _____ your suggestion into consideration.

3. Mary was not able to _____ her parents' attention as they were always busy.

4. Nigeria _____ independence from Great Britain in 1960.

5. Patrick _____ control of the business after his father died and turned it into a totally successful enterprise.

6. My best friend tried to _____ admission to Indiana University of Pennsylvania, but he did not succeed.

7. The U.S. government may have to _____ a high tax on tobacco and gasoline in order to meet its budget goals.

8. He tried to _____ access to my computer files.

9. Where did you _____ this information?

10. No sooner does John _____ his salary than he spends it all; then he has to borrow money for the next month's expenses.

11. John has been trying to _____ a solution to his financial problem.

12. The United States has been trying to _____ intelligence on Iraq's intentions in the Persian Gulf for a long time.

13. Jorge _____ many friends soon after he arrived at Indiana, Pa.; now he knows many people.

14. The scared child _____ her mother's hand.

15. You would be well advised to _____ heed of what he has to say.

16. He _____ an honor for his distinguished work in chemistry.

17. I will strive to _____ my ideals.

18. I can _____ no justification for your behavior.

19. I _____ an apology for that terrible insult you hurled at me in public.

20. Bacon is known to have said: "In _____ revenge, a man is but equal to his enemy, but in passing it over he is his superior."

21. Diligence is one good means of _____ success.

22. In a movie I saw recently, a father tries to _____ vengeance on his son's killers.

23. She tried to _____ Bill's affection by cooking for him.

24. Some Sikh separatists in Cashmere are _____ autonomy.

25. I am gradually _____ competence.

26. It was not easy, but I finally succeeded in _____ his confidence.

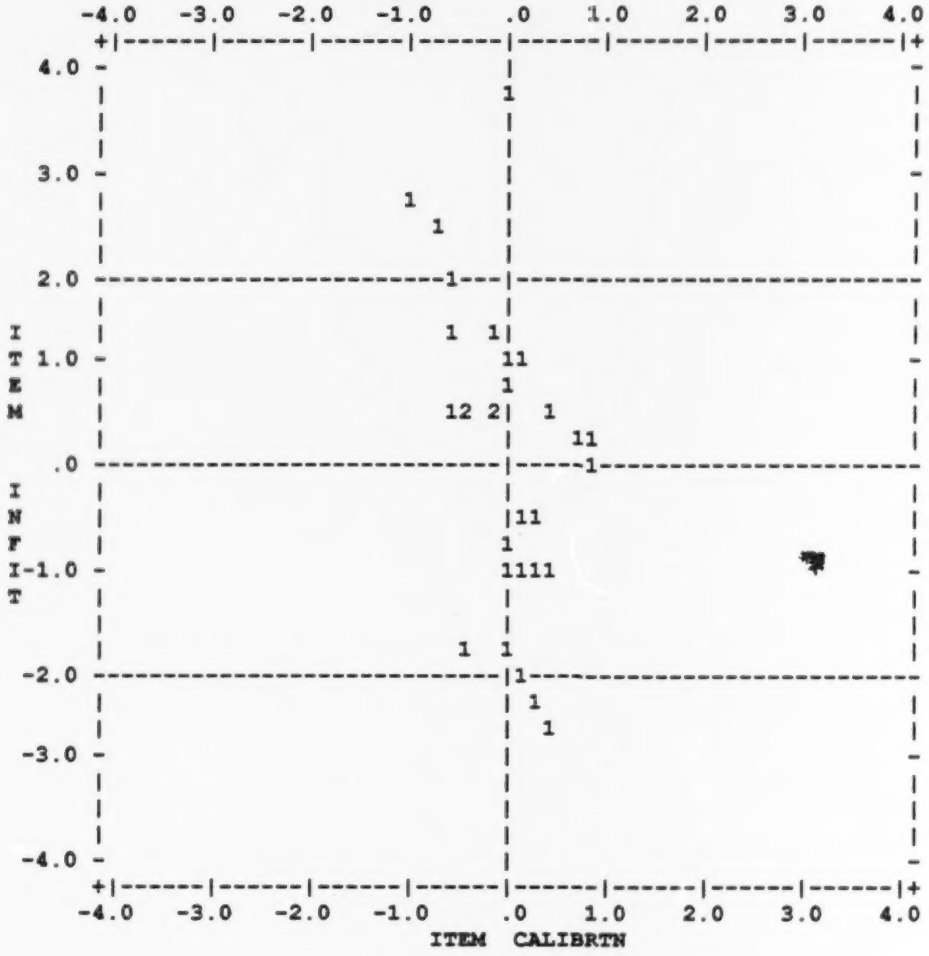
27. She _____ great satisfaction from doing research.

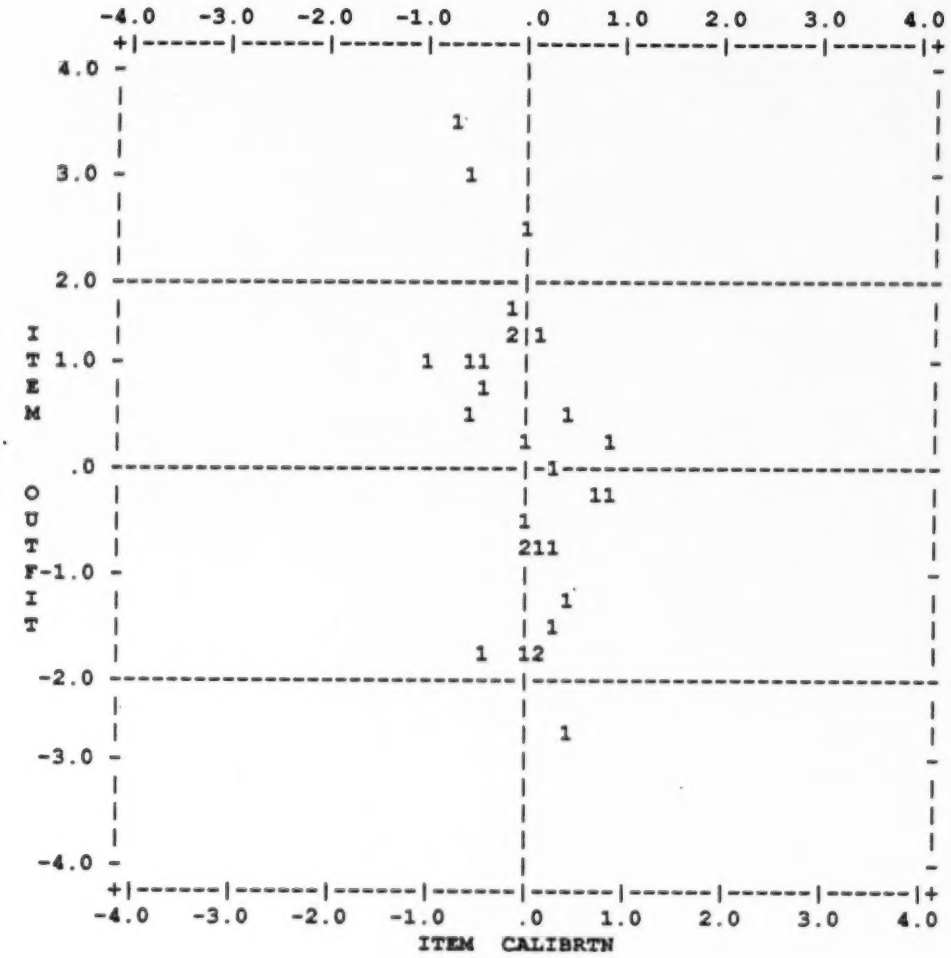
28. I do not understand where he _____ the time to do so much research on top of all his responsibilities at work; he mustn't get more than two hours of sleep each day.

29. The thieves _____ entry to the house through an unlatched window.

30. She worked hard and soon _____ the rank of captain.

Appendix IV. Plots of item infits and outfits




-4.0 -3.0 -2.0 -1.0 .0 1.0 2.0 3.0 4.0
4.0
3.0
2.0
1.0
.0
-1.0
-2.0
-3.0
-4.0
ITEM OUTFIT
ITEM CALIBRTN